

10/539047

JC17 Rec'd PCT/PTO 15 JUN 2005

## AMENDMENT

(Amendment by Regulation of Law Article 11)

To : Examiner of the Patent Office

(seal)

PCT

November 20, 2003

Received

1. Identification of the International Application

PCT/JP03/01434

2. Applicant

Name: NAKADAI, Kazuhiro

Address: 86, Usui, Sakura-shi,  
Chiba 285-0863 JAPAN

Country of nationality: JAPAN

Country of residence: JAPAN

3. AGENT

Name: HIRAYAMA, Kazuyuki

Patent Attorney (8287)

Address: 6th Floor, Shinjukugyoen Bldg.  
3-10, Shinjuku 2-chome, Shinjuku-ku,  
Tokyo 160-0022 JAPAN

4. Object of Amendment

Specification and Claims

5. Content of Amendment

(1) In Specification, p.3, Line 10 (English Translation p.3, Lines 9 – 10) “consisting of the words and their directions which each speaker spoke ” is deleted.

(2) In Specification, p.3, Lines 12 – 13 (English Translation p.3, Lines 13 – 14) “obtained in accordance with acoustic models by said speech recognition process” is amended to “obtained by sound process”.

(3) In Specification, p.3, Lines 14 – 16 (English Translation p.3, Lines 15 – 19) “recognizes the words spoken by respective speakers simultaneously. Said selector may be so constituted as to select said speech recognition process results by majority rule, and provided with a dialogue part to output the speech recognition process results selected by said selector.” is amended to

“acoustic models are provided in each direction so as to respond each speaker and each direction in order to respond the case where a plurality of speakers speak to said robot from different direction with the robot’s front direction as a base, said speech recognition engine executes said speech recognition process in parallel by using each of said acoustic models separately for said one sound source obtained from the sound source separation. Said selector is so constituted as to calculate cost function values based on the recognition result by the speech recognition process and the speaker’s direction, upon integration of the speech recognition process result, and to judge the speech recognition process result having the maximum value of the cost function as the most reliable speech recognition result. Further, and provided with a dialogue part to output the speech recognition process results selected by

said selector.”

(4) In Specification, p.4, Line 8 (English Translation p.4, Line 13) “wherein the auditory module” is amended to “wherein the auditory module is provided with acoustic models for each speaker, and for each direction to respond each direction in order to respond the case where a plurality of speakers speak to said robot from different direction with the robot’s front direction as a base, wherein the auditory module”.

(5) In Specification, p.4, Lines 13 – 14 (English Translation p.4, Lines 21 – 22) “conducts speech recognition of the sound signals separated from sound source separation using a plurality of acoustic models” is amended to “conducts speech recognition in parallel for said one sound signal separated by sound source separation using each of a plurality of acoustic models separately”.

(6) In Specification, p.6, Line 19 (English Translation p.6, Line 33) “wherein the auditory module” is amended to “wherein the auditory module is provided with acoustic models for each speaker, and for each direction to respond each direction, in order to respond the case where a plurality of speakers speak to said robot from different direction with the robot’s front direction as a base, wherein the auditory module”.

(7) In Specification, p.6, Lines 24 – 25 (English Translation p.7, Lines 4 – 6) “conducts speech recognition of the sound signals separated by sound sources separation using a plurality of acoustic models,” is amended to “conducts speech recognition in parallel for

said one sound signal separated by sound source separation using each of a plurality of acoustic models separately,”.

(8) In claims, p.31, Line 1 of Claim 1 (English Translation p.33, Lines 2 – 3 of Claim 1) “consisting of the words spoken by each speaker and their directions combined”, and also on Lines 3 – 4 (English Translation, Line 8 of Claim 1) “for each of said acoustic models” are deleted, and also,

on Line 6 (English Translation, Lines 10 – 11 of Claim 1) “the words spoken simultaneously by each speaker are each recognized” is amended to “, in order to respond the case where a plurality of speakers speak to said robot from different directions with the robot’s front direction as the base, sound models are provided in each direction so to respond each speaker, and each direction, wherein said speech recognition engine uses each of said acoustic models separately for one sound signal separated by sound source separation, and executes said speech recognition process in parallel”.

(9) In claims, p.31, Lines 1 – 2 of Claim 2 (English Translation p.33, Lines 2 – 3 of Claim 2) “is made up as to select the speech recognition process results by majority vote” is amended to “calculates the cost function value, upon integrating the speech recognition process result, based on the recognition result by the speech recognition process and the speaker’s direction, and judges the speech recognition process result having the maximum value of the cost function as the most reliable speech recognition result”.

(10) In claims, p.32 – 33, Claims 4 and 5 (English Translation p.33

– 34, Claims 4 and 5) “wherein the auditory module collects” is amended to “in order for the auditory module to respond the case where a plurality of speakers speak to said robot from different directions with the robot’s front direction as the base, acoustic models are provided in each direction so to respond each speaker, and each direction, wherein the auditory module collects”, and also “conducts speech recognition of separated sound signals from respective sound sources using a plurality of acoustic models” is amended to “conducts speech recognition in parallel for one sound signal separated by sound source separation using a plurality of the acoustic models”.

(11) In claims p.34 (English Translation p.36) Claim 12 is added as shown below.

“12. A robotics visual and auditory system as set forth in Claim 4 or Claim 5, wherein the selector calculates the cost function value, upon integrating the speech recognition result, based on the recognition result by the speech recognition and the direction determined by the association module, and judges the speech recognition process result having the maximum value of the cost function as the most reliable speech recognition result.”

(12) In claims p.34 (English Translation p.36) Claim 13 is added as shown below.

“13. A robotics visual and auditory system as set forth in Claim 4 or Claim 5, characterized in that; it recognizes the speaker’s name based on the acoustic model utilized to obtain speech recognition result.”

**6. List of Attached Documents**

(1) Specification p.3, 3/1, 4, 4/1, 6, 6/1 (English Translation p.3, 3/1, 4, 4/1, 6,6/1, 7)

(2) claims p.31, 32, 33, 33/1, 34 (English Translation p. 33, 33/1, 34, 34/1, 35, 35/1, 36, 36/1)

Disclosure of the Invention

[0008] It is the objective of the present invention, taking into consideration the above-mentioned problems, to provide a robot audiovisual system capable of recognition of sounds separated from respective sound sources. In order to achieve the above-mentioned objective, a first aspect of the robot audiovisual system of the present invention is characterized in that it is provided with a plurality of acoustic models, a speech recognition engine performing speech recognition process to the sound signals separated from respective sound sources, and the selector to integrate a plurality of the speech recognition process results obtained by sound process, and to select any one of the speech recognition process results, thereby acoustic models are provided in each direction so as to respond each speaker and each direction in order to respond the case where a plurality of speakers speak to said robot from different direction with the robot's front direction as a base, said speech recognition engine executes said speech recognition process in parallel by using each of said acoustic models separately for said one sound source obtained from the sound source separation. Said selector is so constituted as to calculate cost function values based on the recognition result by the speech recognition process and the speaker's direction, upon integration of the speech recognition process result, and to judge the speech recognition process result having the maximum value of the cost function as the most reliable speech recognition result. Further, and provided with a dialogue part to output the speech recognition process results selected by said selector.

[0009] According to said first aspect, by using a plurality of acoustic models based on the sound signals conducted sound source localization and sound source separation, the speech recognition processes are performed, respectively, and, by integrating by the selector the speech recognition process results, the most reliable speech recognition result is judged.

[0010] In order also to achieve the above-mentioned objective, a

second aspect of the robotics visual and auditory system of the present invention is provided with an auditory module which is provided at least with a pair of microphones to collect external sounds, and, based on sound signals from the microphones, determines a direction of at least one speaker by sound source separation and localization by grouping based on pitch extraction and harmonic sounds, a face module which is provided a camera to take images of a robot's front, identifies each speaker, and extracts his face event from each speaker's face recognition and

localization, based on images taken by the camera, a motor control module which is provided with a drive motor to rotate the robot in the horizontal direction, and extracts motor event, based on a rotational position of the drive motor, an association module which determines each speaker's direction, based on directional information of sound source localization of the auditory event and face localization of the face event, from said auditory, face, and motor events, generates an auditory stream and a face stream by connecting said events in the temporal direction using a Kalman filter for determinations, and further generates an association stream associating these streams, and an attention control module which conduct an attention control based on said streams, and drive-controls the motor based on an action planning results accompanying the attention control, wherein the auditory module is provided with acoustic models for each speaker, and for each direction to respond each direction in order to respond the case where a plurality of speakers speak to said robot from different direction with the robot's front direction as a base, wherein the auditory module collects sub-bands having interaural phase difference (IPD) or interaural intensity difference (IID) within a predetermined range by an active direction pass filter having a pass range which, according to auditory characteristics, becomes minimum in the frontal direction, and larger as the angle becomes wider to the left and right, based on an accurate sound source directional information from the association module, and conducts sound source separation by restructuring a wave shape of a sound source, conducts speech recognition in parallel for said one sound signal separated by sound source separation using each of a plurality of acoustic models separately, integrates speech recognition results from each acoustic model by a selector, and judges the most reliable speech recognition result among the speech recognition results.

[0011] According to such second aspect, the auditory module conducts pitch extraction utilizing harmonic sound from the sound from the outside target collected by the microphone, thereby obtains the direction of each sound source, identifies individual speakers, and extracts said auditory event. The face module extracts individual speakers' face events by face

recognition and localization of each speaker by pattern recognition from the images photographed by the camera. Further, the motor control module extracts motor event by detecting the robot's direction based on the rotating position of the drive motor which rotates the robot

separation using a plurality of acoustic models, as described above, and integrates the speech recognition result by each acoustic model by the selector, and judges the most reliable speech recognition result. Thereby, accurate speech recognition in real time and real environment is possible by using a plurality of acoustic models, compared with conventional speech recognition, as well as speech recognition result is integrated by the selector by each acoustic model, the most reliable speech recognition result is judged, thereby more accurate speech recognition is possible.

[0017] In order also to achieve the above-mentioned objective, a third aspect of the robotics visual and auditory system of the present invention is provided with an auditory module which is provided at least with a pair of microphones to collect external sounds, and, based on sound signals from the microphones, determines a direction of at least one speaker by sound source separation and localization by grouping based on pitch extraction and harmonic sounds, a face module which is provided a camera to take images of a robot's front, identifies each speaker, and extracts his face event from each speaker's face recognition and localization, based on images taken by the camera, a stereo module which extracts and localizes a longitudinally long matter, based on a parallax extracted from images taken by a stereo camera, and extracts stereo event, a motor control module which is provided with a drive motor to rotate the robot in the horizontal direction, and extracts motor event, based on a rotational position of the drive motor, an association module which determines each speaker's direction, based on directional information of sound source localization of the auditory event and face localization of the face event, from said auditory, face, stereo, and motor events, generates an auditory stream, a face stream and a stereo visual stream by connecting said events in the temporal direction using a Kalman filter for determinations, and further generates an association stream associating these streams, and an attention control module which conduct an attention control based on said streams, and drive-controls the motor based on an action planning results accompanying the attention control, wherein the auditory module is provided with acoustic models for each speaker, and for each direction to respond each direction,

in order to respond the case where a plurality of speakers speak to said robot from different direction with the robot's front direction as a base, wherein the auditory module collects sub-bands having interaural phase difference (IPD) or interaural intensity difference (IID)

within a predetermined range by an active direction pass filter having a pass range which, according to auditory characteristics, becomes minimum in the frontal direction, and larger as the angle becomes wider to the left and right, based on an accurate sound source directional information from the association module, and conducts sound source separation by restructuring a wave shape of a sound source, conducts speech recognition in parallel for said one sound signal separated by sound source separation using each of a plurality of acoustic models separately, integrates speech recognition results from each acoustic model by a selector, and judges the most reliable speech recognition result among the speech recognition results.

[0018] According to such third aspect, the auditory module conducts pitch extraction utilizing harmonic sound from the sound from the outside target collected by the microphone, thereby obtains the direction of each sound source, and extracts the auditory event. The face module extracts individual speakers' face events by identifying each speaker from face recognition and localization of each speaker by pattern recognition from the images photographed by the camera. Further, the stereo module extracts and localizes a longitudinally long matter, based on a parallax extracted from images taken by the stereo camera, and extracts stereo event. Further, the motor control module extracts motor event by detecting the robot's direction based on the rotating position of a drive motor which rotates the robot horizontally.

[0019] In this connection, said event indicates that there are sounds, faces, and longitudinally long matters to be detected at each time, or the state in which the drive motor is rotated, and said stream indicates the event connected temporally continuous with, for example, a Kalman filter or others while correcting errors.

[0020] Here, the association module generates each speaker's auditory, face, and stereo visual streams by determining each speaker's direction from the sound source localization of an auditory event and the face localization of a face event, based on thus extracted auditory, face, stereo, and motor events, and further generates an association stream associating these streams. Here, the association stream gives the image

**What is claimed is:**

1(Amended). A robotics visual and auditory system comprising;

a plurality of acoustic models,

a speech recognition engine for executing speech recognition processes to separated sound signals from respective sound sources by using the acoustic models, and

a selector for integrating a plurality of speech recognition process results obtained by the speech recognition process, and selecting any one of speech recognition process results,

wherein, in order to respond the case where a plurality of speakers speak to said robot from different directions with the robot's front direction as the base, the acoustic models are provided with respect to each speaker and each direction so to respond each direction,

wherein the speech recognition engine uses each of said acoustic models separately for one sound signal separated by sound source separation, and executes said speech recognition process in parallel.

2(Amended). A robotics visual and auditory system as set forth in Claim 1, wherein the selector calculates the cost function value, upon integrating the speech recognition process result, based on the recognition result by the speech recognition process and the speaker's direction, and judges the speech recognition process result having the maximum value of the cost function as the most reliable speech recognition result.

3(original). A robotics visual and auditory system as set forth in Claim 1 or Claim 2, wherein it is provided with a dialogue part to output the speech recognition process results selected by the selector to outside.

4(Amended). A robotics visual and auditory system comprising;

an auditory module which is provided at least with a pair of microphones to collect external sounds, and, based on sound signals from the microphones, determines a direction of at least one speaker by sound

source separation and localization by grouping based on pitch extraction and harmonic sounds,

a face module which is provided a camera to take images of a robot's front, identifies each speaker, and extracts his face event from each speaker's face recognition and localization, based on images taken by the camera,

a motor control module which is provided with a drive motor to rotate the robot in the horizontal direction, and extracts motor event, based on a

rotational position of the drive motor,

an association module which determines each speaker's direction, based on directional information of sound source localization of the auditory event and face localization of the face event, from said auditory, face, and motor events, generates an auditory stream and a face stream by connecting said events in the temporal direction using a Kalman filter for determinations, and further generates an association stream associating these streams, and

an attention control module which conduct an attention control based on said streams, and drive-controls the motor based on an action planning results accompanying the attention control,

in order for the auditory module to respond the case where a plurality of speakers speak to said robot from different directions with the robot's front direction as the base, acoustic models are provided in each direction so to respond each speaker, and each direction,

wherein the auditory module collects sub-bands having interaural phase difference (IPD) or interaural intensity difference (IID) within a predetermined range by an active direction pass filter having a pass range which, according to auditory characteristics, becomes minimum in the frontal direction, and larger as the angle becomes wider to the left and right, based on an accurate sound source directional information from the association module, and conducts sound source separation by restructuring a wave shape of a sound source, conducts speech recognition in parallel for one sound signal separated by sound source separation using a plurality of the acoustic models, integrates speech recognition results from each acoustic model by a selector, and judges the most reliable speech recognition result among the speech recognition results.

5(Amended). A robotics visual and auditory system comprising;

an auditory module which is provided at least with a pair of microphones to collect external sounds, and, based on sound signals from the microphones, determines a direction of at least one speaker

by sound source separation and localization by grouping based on pitch extraction and harmonic sounds,

a face module which is provided a camera to take images of a robot's front, identifies each speaker, and extracts his face event from each speaker's face recognition and localization, based on images taken by the camera,

a stereo module which extracts and localizes a longitudinally long matter, based on a parallax extracted from images taken by a stereo camera, and extracts stereo event,

a motor control module which is provided with a drive motor to rotate the robot in the horizontal direction, and extracts motor event, based on a rotational position of the drive motor,

an association module which determines each speaker's direction, based on directional information of sound source localization of the auditory event and face localization of the face event, from said auditory, face, stereo, and motor events, generates an auditory stream, a face stream and a stereo visual stream by connecting said events in the temporal direction using a Kalman filter for determinations, and further generates an association stream associating these streams, and

an attention control module which conduct an attention control based on said streams, and drive-controls the motor based on an action planning results accompanying the attention control,

in order for the auditory module to respond the case where a plurality of speakers speak to said robot from different directions with the robot's front direction as the base, acoustic models are provided in each direction so to respond each speaker, and each direction,

wherein the auditory module collects sub-bands having interaural phase difference (IPD) or interaural intensity difference (IID) within a predetermined range by an active direction pass filter having a pass range which, according to auditory characteristics, becomes minimum in the frontal direction, and larger as the angle becomes wider to the left and right, based on an accurate sound source directional information from the association module, and conducts sound source separation by restructuring a wave shape of a sound source, conducts speech recognition in parallel for one sound signal separated by sound source separation using a plurality of the acoustic models, integrates speech recognition results from each acoustic model by a selector, and judges the most reliable speech recognition result among the speech recognition results.

6(original). A robotics visual and auditory system as set forth in Claim 4 or Claim 5, characterized in that;

when the speech recognition by the auditory module failed, the attention control module is made up as to collect speeches again from the microphones with the microphones and the camera turned to the sound

source direction of the sound signals, and to perform again speech recognition of the speech by the auditory module, based on the sound signals conducted sound source localization and sound source separation.

7(original). A robotics visual and auditory system as set forth in Claim 4 or Claim 5, characterized in that;

the auditory module refers to the face event from the face module upon performing the speech recognition.

8(original). A robotics visual and auditory system as set forth in Claim 5, characterized in that;

the auditory module refers to the stereo event from the stereo module upon performing the speech recognition.

9(original). A robotics visual and auditory system as set forth in Claim 5, characterized in that;

the auditory module refers to the face event from the face module and the stereo event from the stereo module upon performing the speech recognition.

10(original). A robotics visual and auditory system as set forth in Claim 4 or Claim 5, wherein it is provided with a dialogue part to output the speech recognition results judged by the auditory module to outside.

11(original). A robotics visual and auditory system as set forth in Claim 4 or Claim 5, wherein a pass range of the active direction pass filter can be controlled for each frequency.

12(New). A robotics visual and auditory system as set forth in Claim 4 or Claim 5, wherein the selector calculates the cost function value, upon integrating the speech recognition result, based on the recognition result by the speech recognition and the direction determined by the association module, and judges the speech recognition process result having the maximum value of the cost function as the most reliable speech

recognition result.

13(New). A robotics visual and auditory system as set forth in Claim 4 or Claim 5, characterized in that; it recognizes the speaker's name based on the acoustic model utilized to obtain speech recognition result.